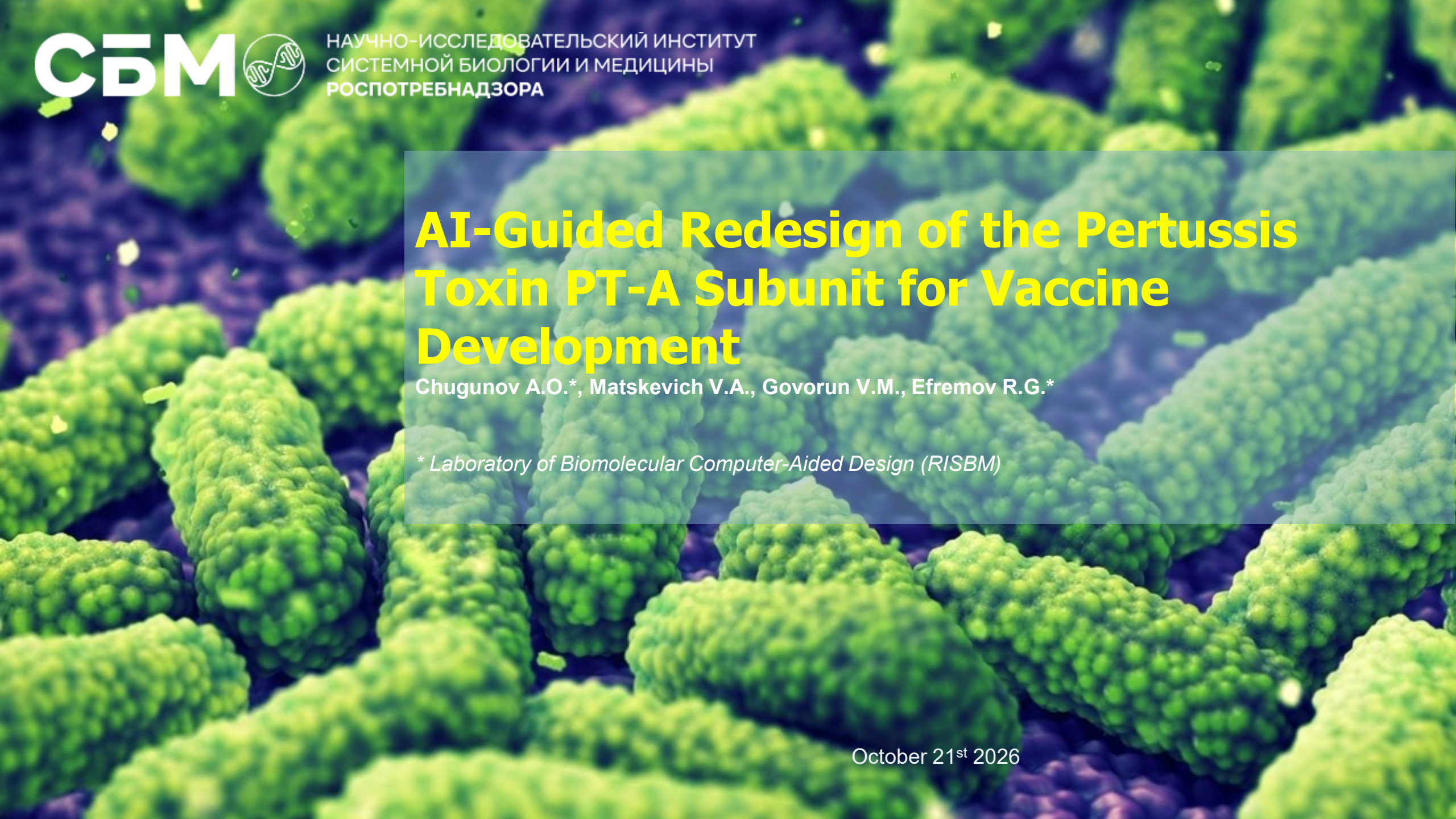


НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
СИСТЕМНОЙ БИОЛОГИИ И МЕДИЦИНЫ
РОСПОТРЕБНАДЗОРА

AI-Guided Redesign of the Pertussis Toxin PT-A Subunit for Vaccine Development

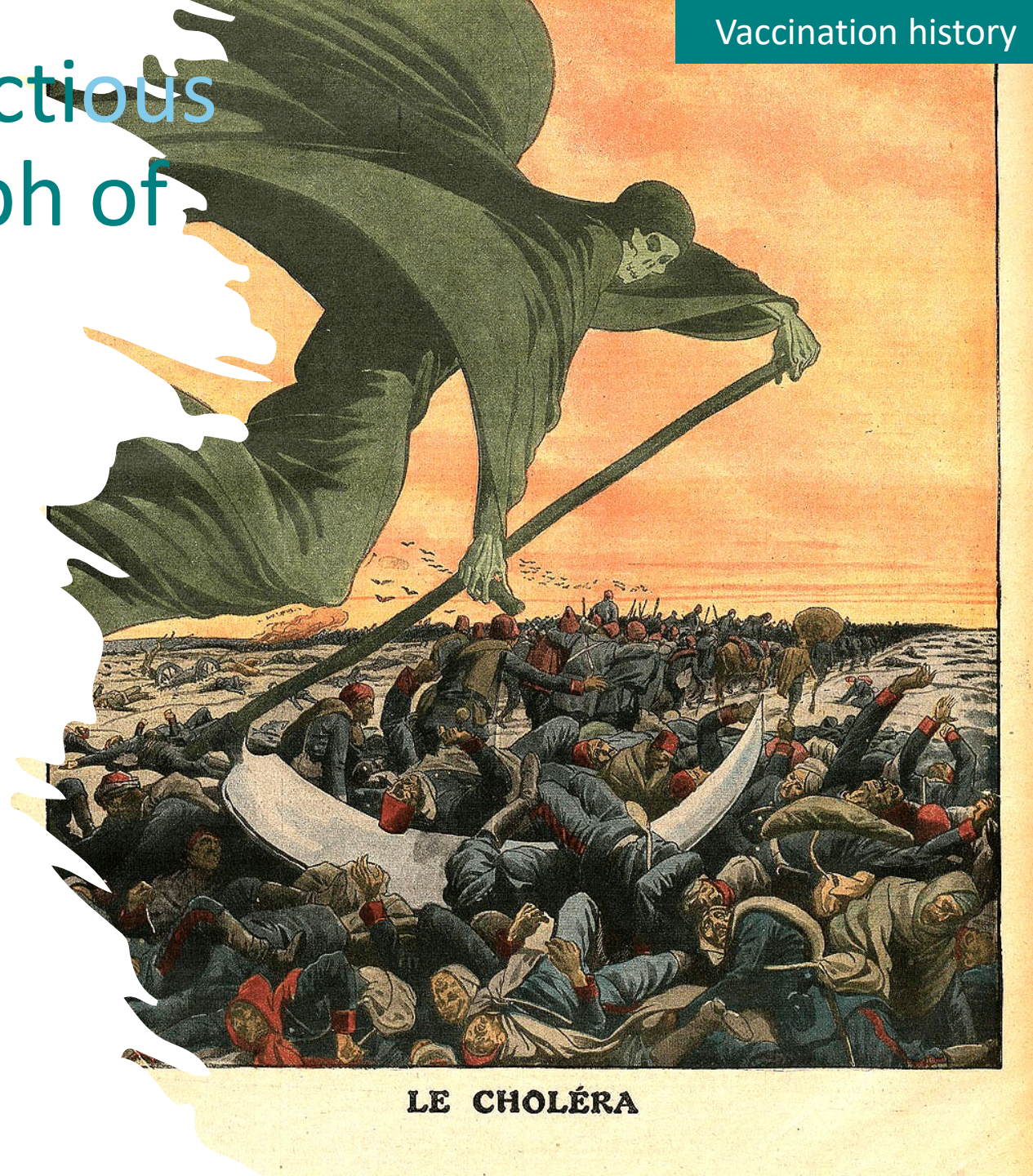
Chugunov A.O.*, Matskevich V.A., Govorun V.M., Efremov R.G.*

** Laboratory of Biomolecular Computer-Aided Design (RISBM)*

October 21st 2026

Low mortality from infectious diseases today: a triumph of hygiene and medicine

- Plague, smallpox, and cholera — major killers in centuries past
  - Only about 3,000 plague cases reported worldwide in 2010–2015
  - Smallpox eradicated worldwide (1980)
- In the Middle Ages, disease was attributed to “miasmas”
- One defense: quarantine (just ask Romeo)
- Modern advances:
  - Hygiene
  - Disinfection
  - Antibiotics, other medicines, and vaccines



Pertussis — the cough that kills

- Before the 1950s, this disease killed three times as many children as measles, mumps, chickenpox, rubella, scarlet fever, diphtheria, polio, and meningitis combined.
  - Early symptoms resemble a common cold
  - Violent coughing fits followed by a high-pitched "whoop" (French *coq* — rooster)
  - Face turns blue and puffy
  - Eyes become bloodshot
  - Tongue protrudes as far as possible
- The causative agent, *Bordetella pertussis*, is an aerobic, Gram-negative coccobacillus. It targets the bronchial epithelium and infects only humans.
- What exactly causes the cough remains unclear!

Whooping cough

Pertussis Research History

XV век

Бахадол Рази впервые сообщил о двух эпидемиях коклюша в Персии



1578 год

Г. де Байю описал эпидемию коклюша в Европе



1679 год

Т. Сиденхам дал болезни название «pertussis» (сильный кашель)



1784 год

Коклюш впервые описан в России Н. Амбодик-Максимовичем в «Искусстве повивания»



1847 год

С.Хотовицкий описал развитие болезни и ее симптомы в «Педиятрике»



1906 год

Выделение коклюшных бактерий Ж.Борде и О. Жангу



1938 год

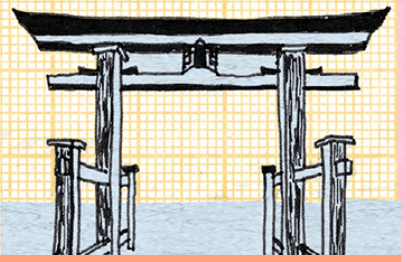
Создание вакцины Кендрик



DTwP precusor (АКДС)

1981 год

Первая бесклеточная вакцина от коклюша появилась в Японии



Pentaxim precusor

2003 год

Секвенирован полный геном коклюшных бактерий



2006 год

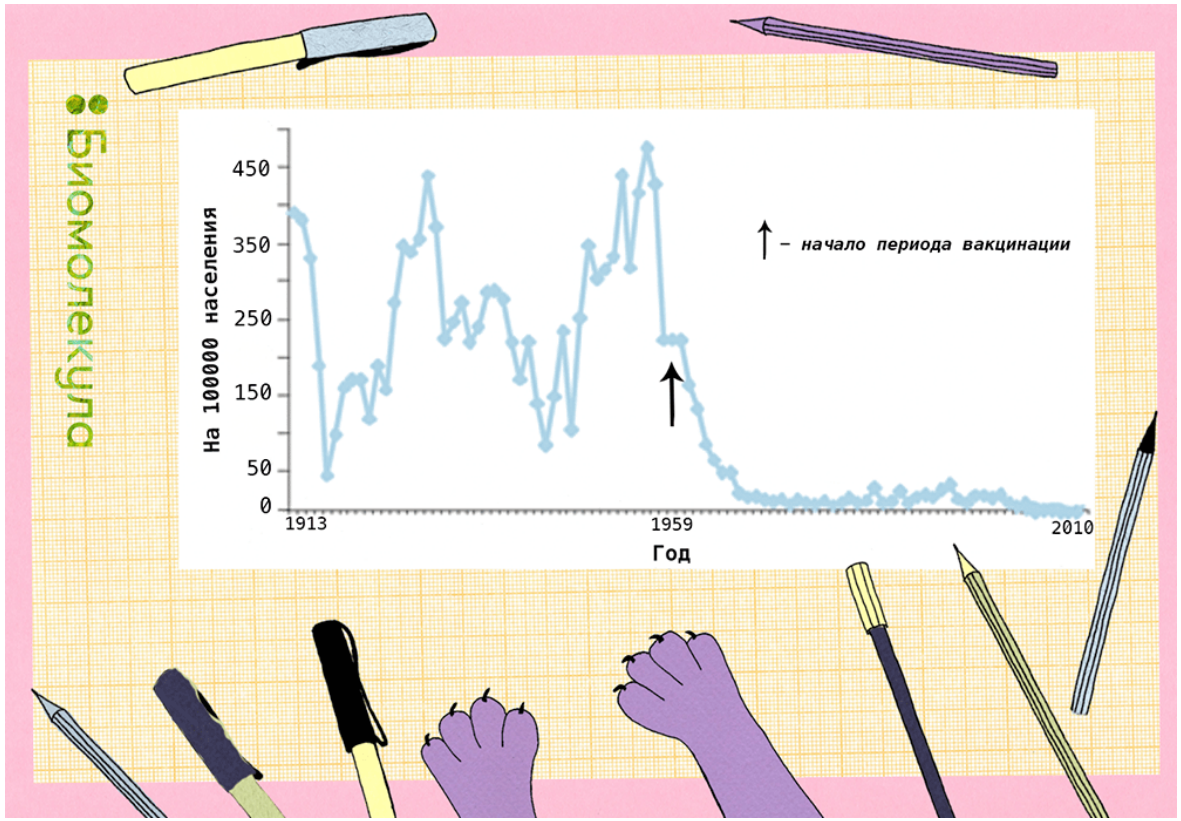
На рынке появилась первая вакцина для взрослых



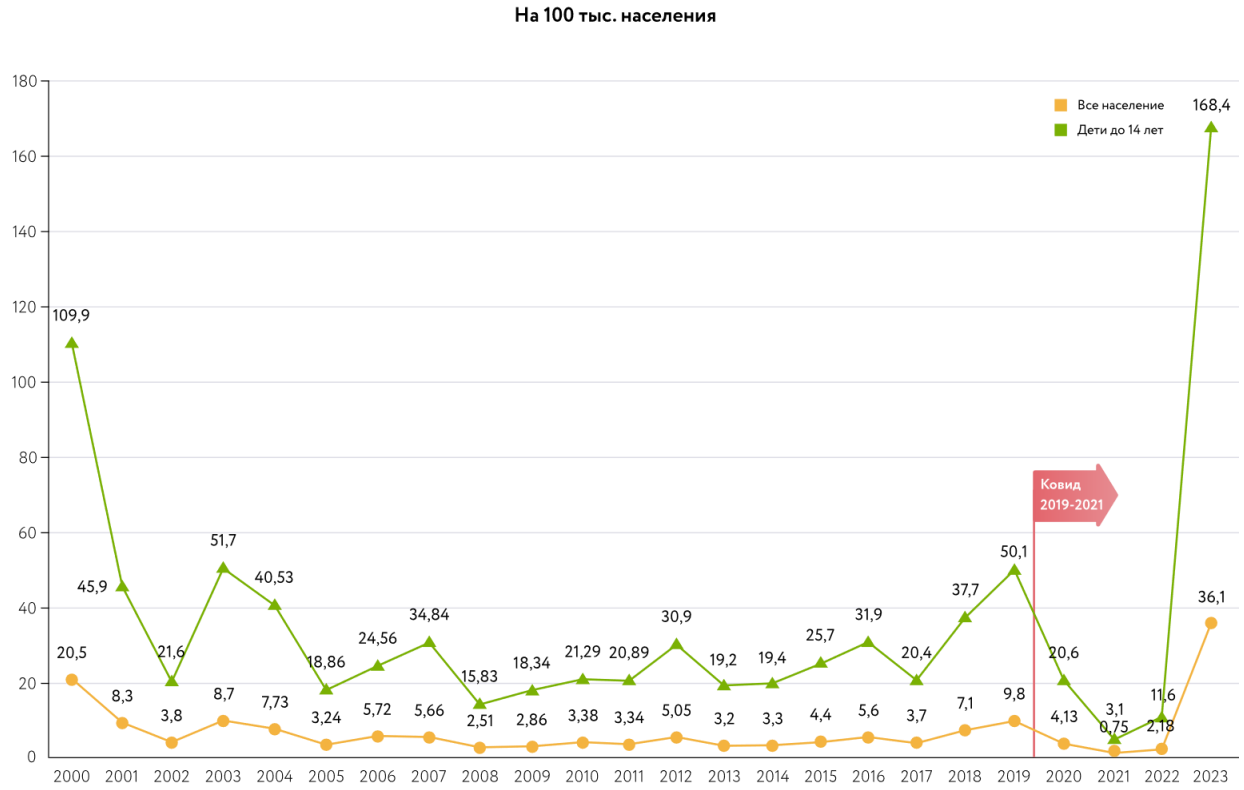
Биомолекула

: Биомолекула: «Коклюш: кашель, который убивает»

Pertussis in Russia: a Resurgence



The Soviet era: the triumph of the DTwP vaccine (АКДС)

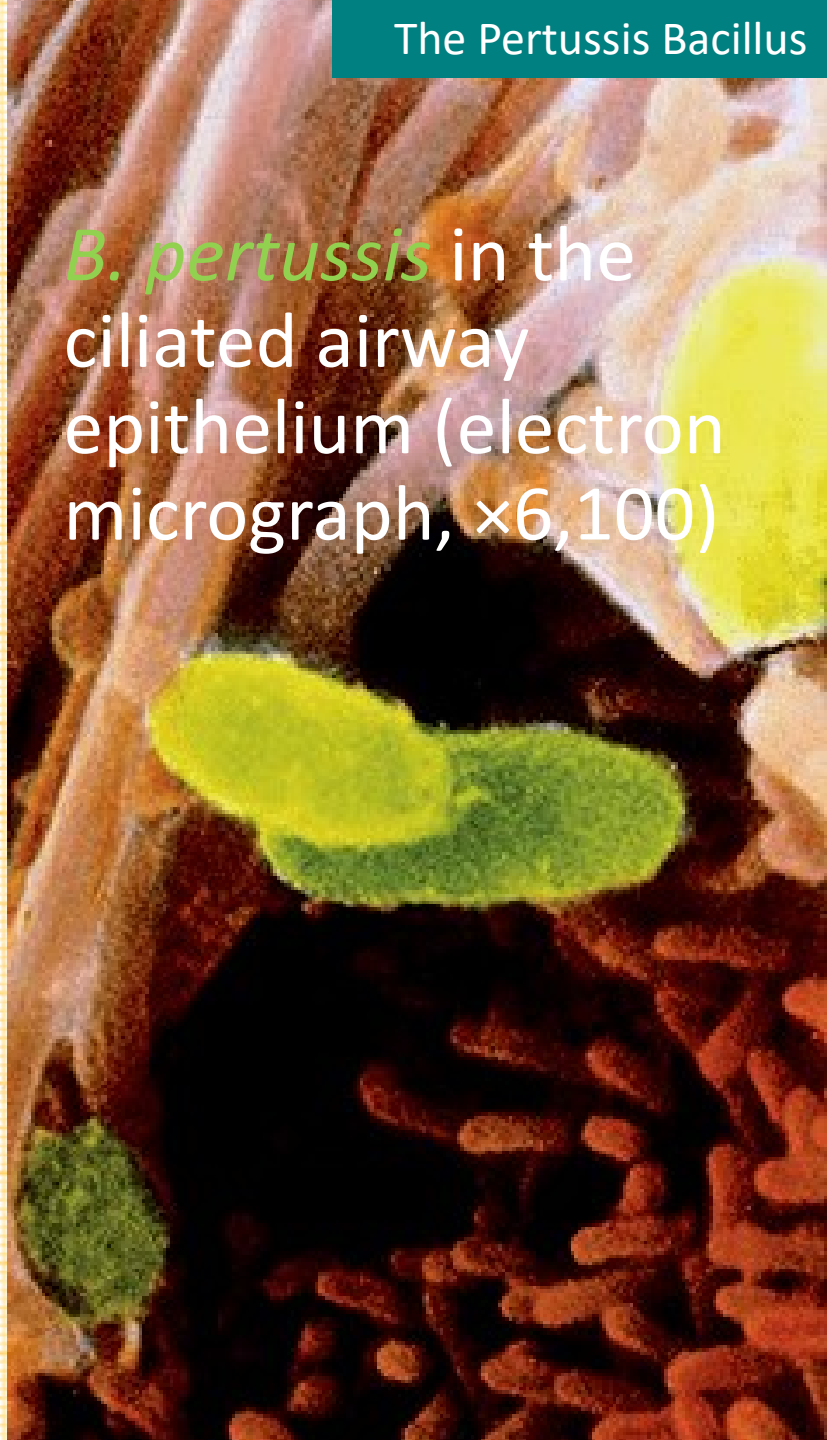


Today: the return of pertussis

- *Sanitary and Epidemiological Well-Being of the Population in the Russian Federation.* State reports, 2000–2022. Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing (Rospotrebnadzor).
- Basov A.A. *Pertussis in 2023–2024: Causes of Rising Incidence and Strategies to Address It.* 16th Annual All-Russian Congress on Infectious Diseases in Honor of Academician V.I. Pokrovsky. March 25, 2024.

The Pertussis Bacillus

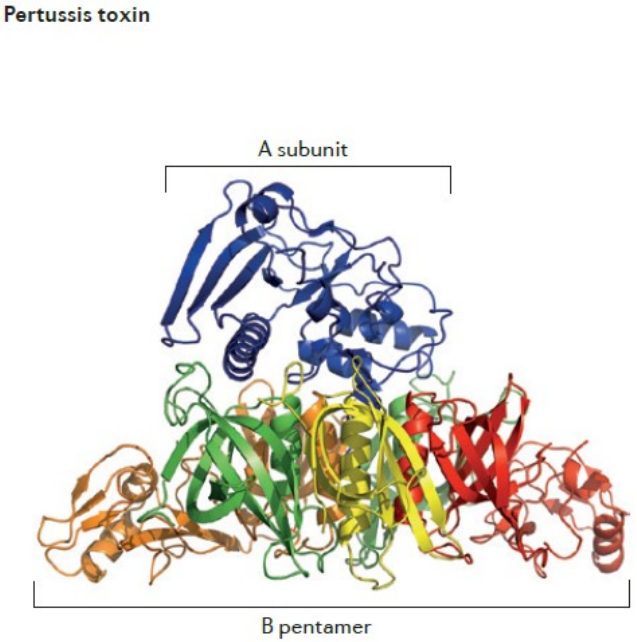


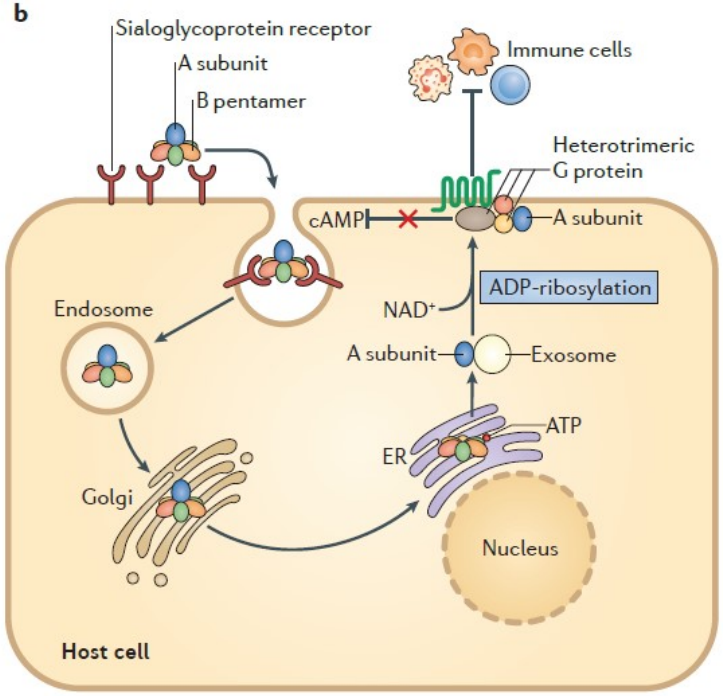


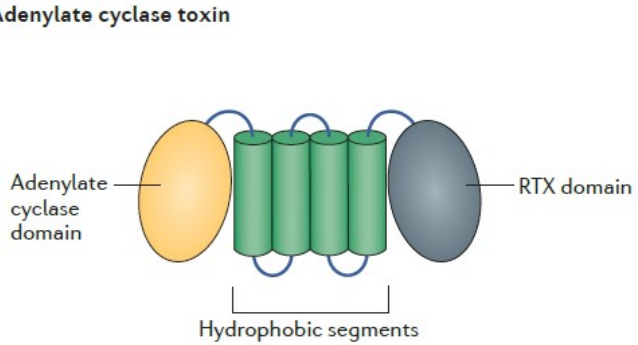
B. pertussis in the ciliated airway epithelium (electron micrograph, ×6,100)

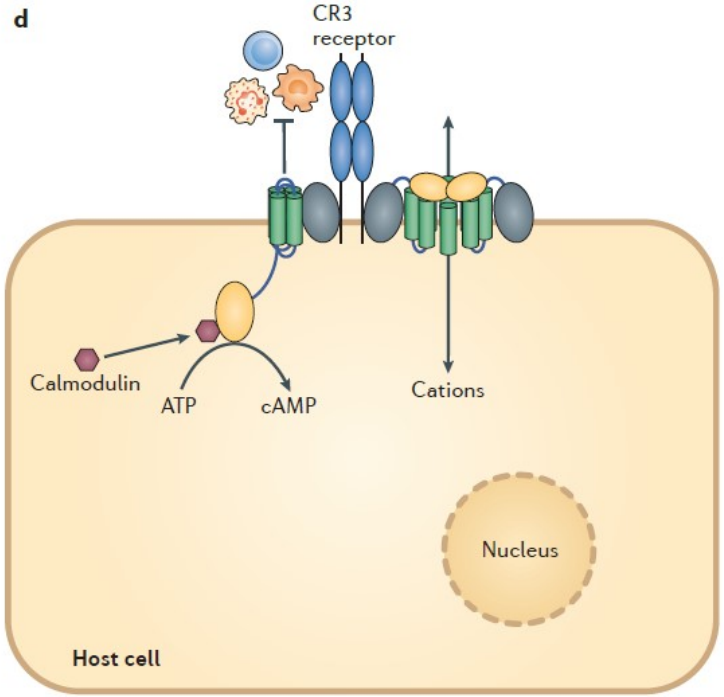
B. Pertussis toxins

- After escaping the endosome, PT inactivates Gα-subunits of G proteins by ADP-ribosylation
- This suppresses cytokine production and chemotaxis, inhibiting cell-mediated immunity
- ACT produces similar effects through a different mechanism



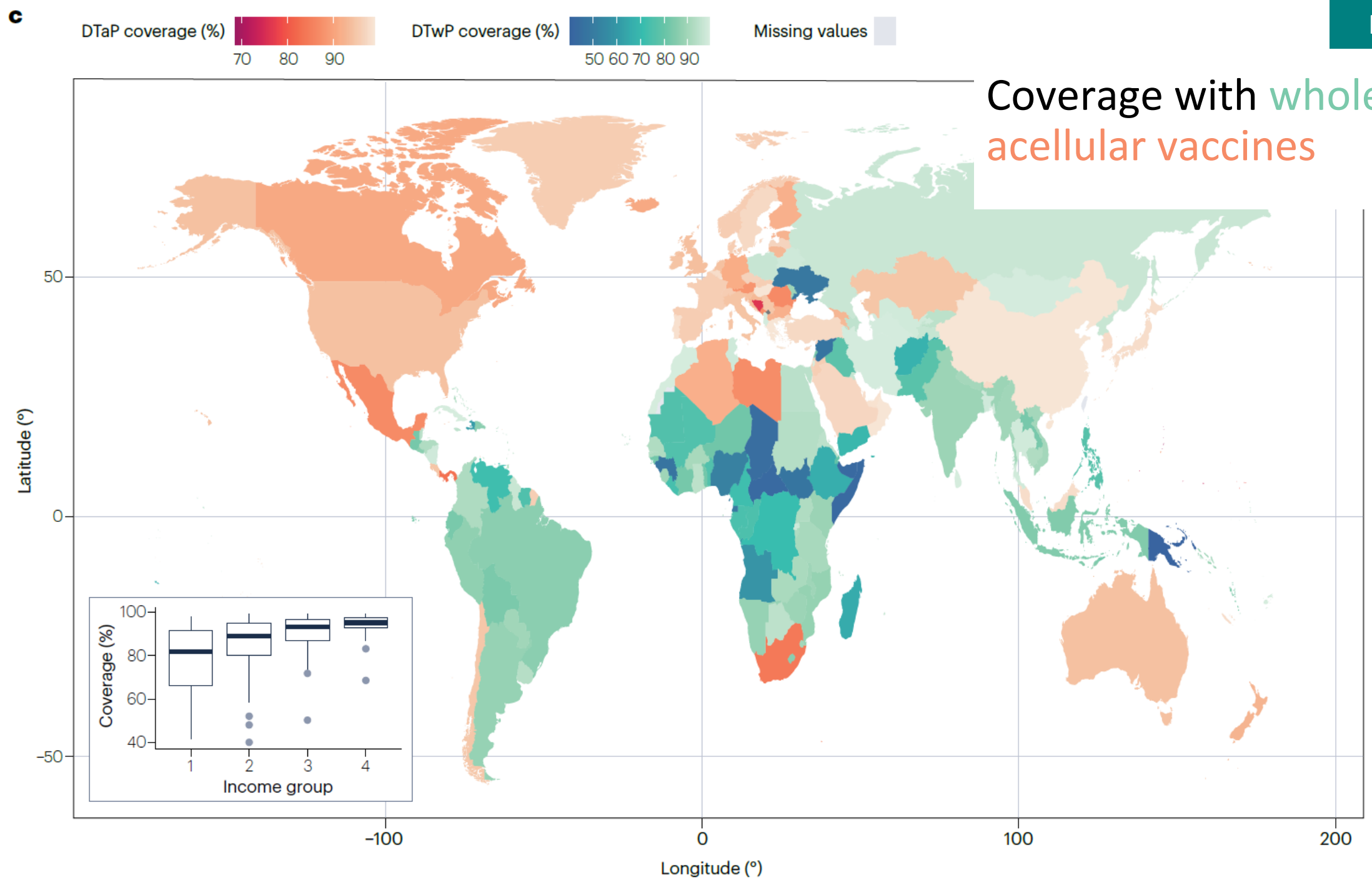






Melvin *et al.*, *Nat. Rev. Microbiol.* 2014

Coverage with whole-cell and acellular vaccines



de Cellès & Rohani, *Nat. Rev. Microbiol.* 2024

Talk outline

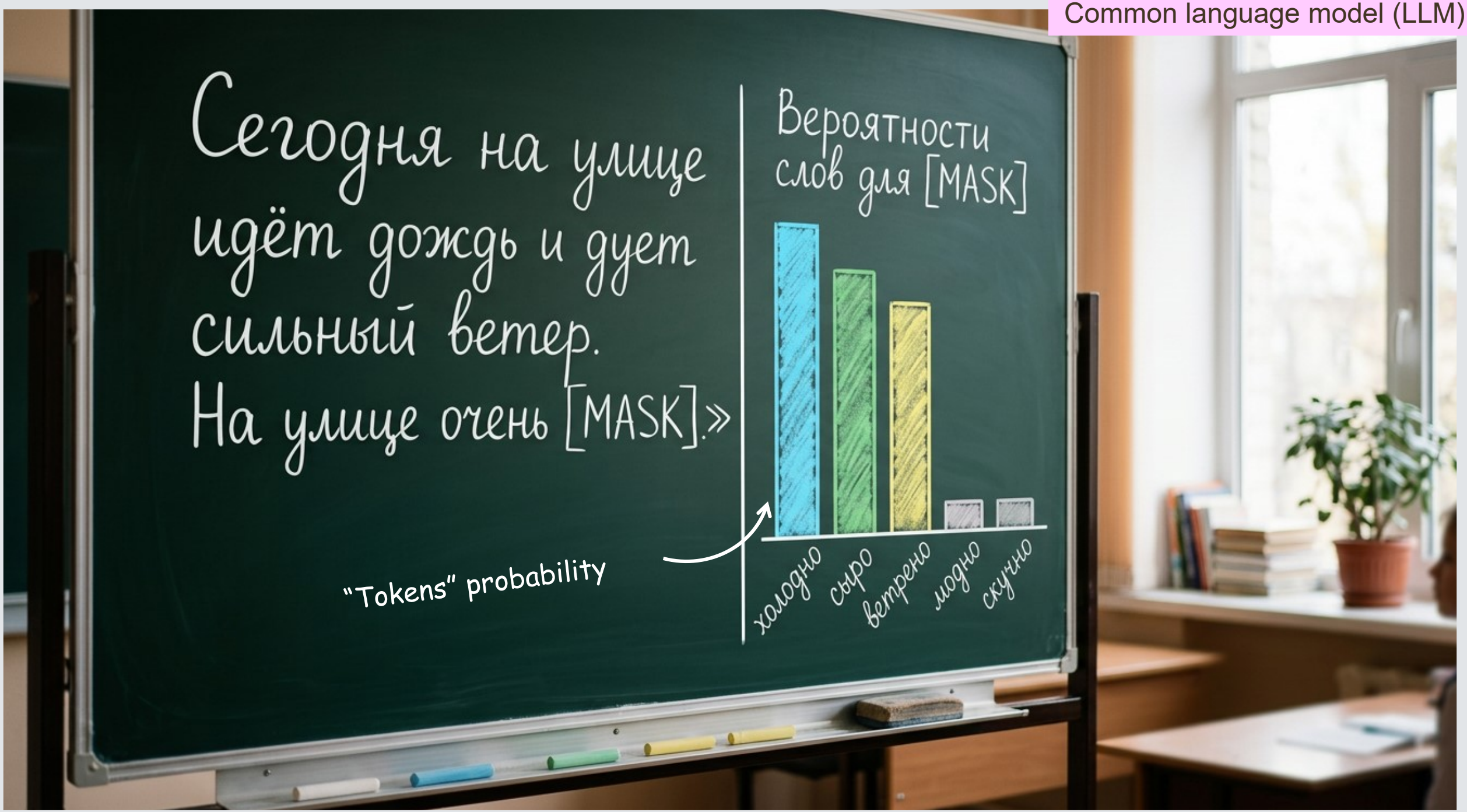
- A brief history of vaccination and pertussis
- Developing a new subunit vaccine
- Protein language models: protein (re)design
- Redesigning PT-A using **AI** and protein engineering
- Initial validation of engineered PT-A

The goal of this work —

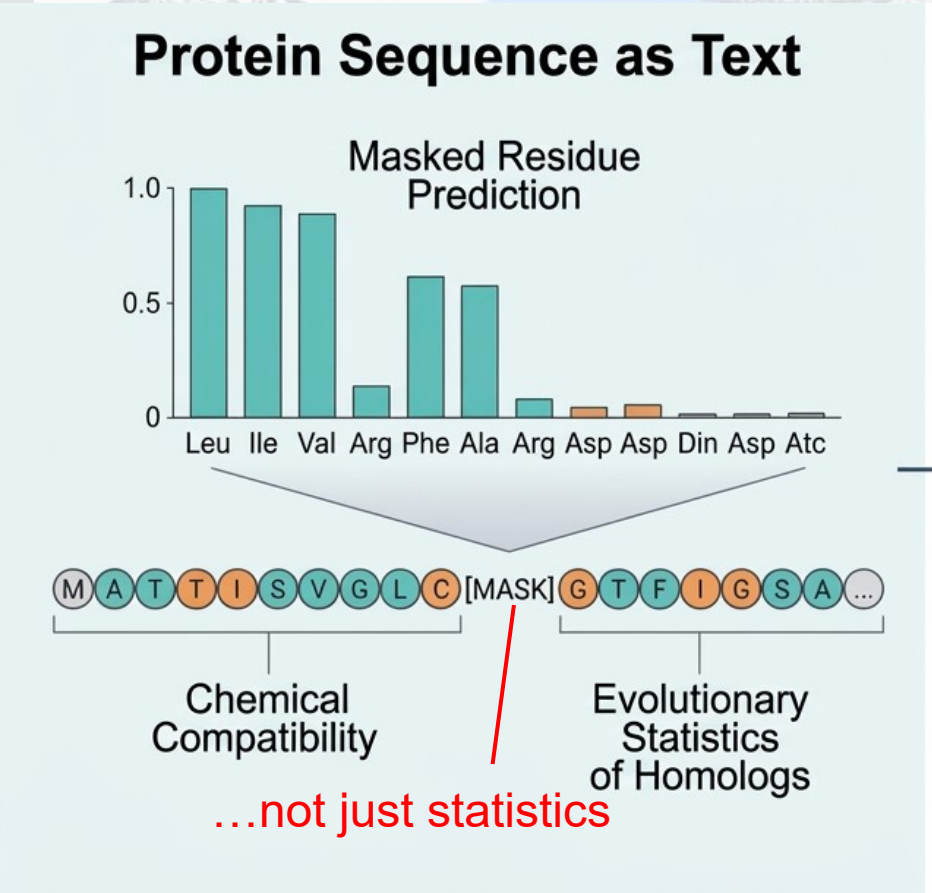
to develop Russia's first subunit vaccine against pertussis (aP)

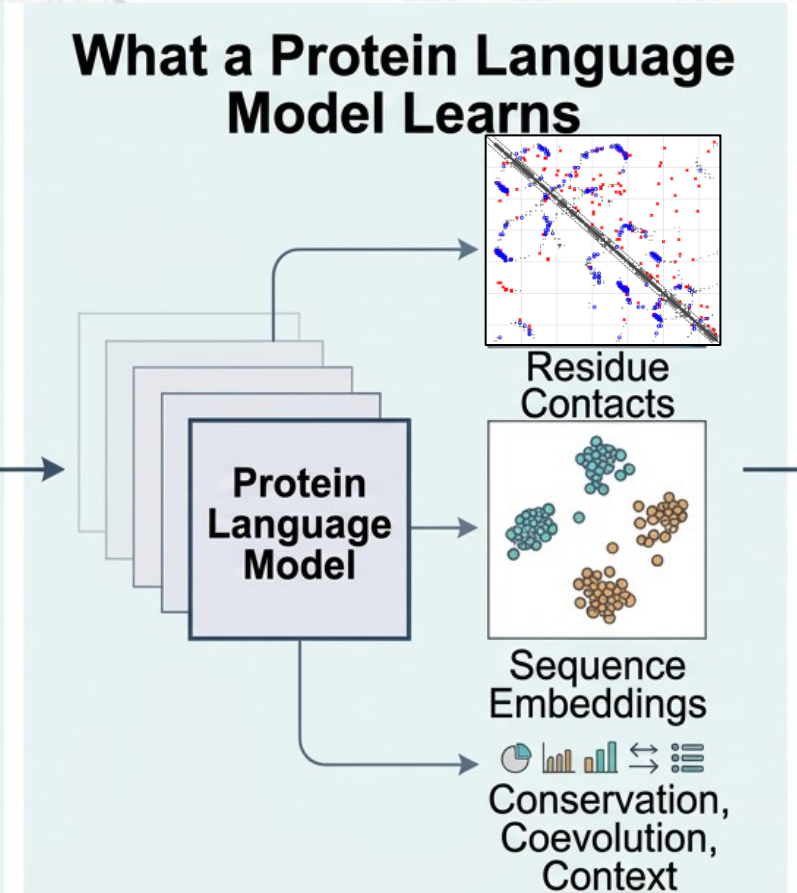
- One component: the gdPT-A toxin
  - Challenge: low recombinant toxin yield, aggregation, and precipitation
  - Solution: redesign PT-A using protein engineering and artificial intelligence methods
- Fusion with other subunits

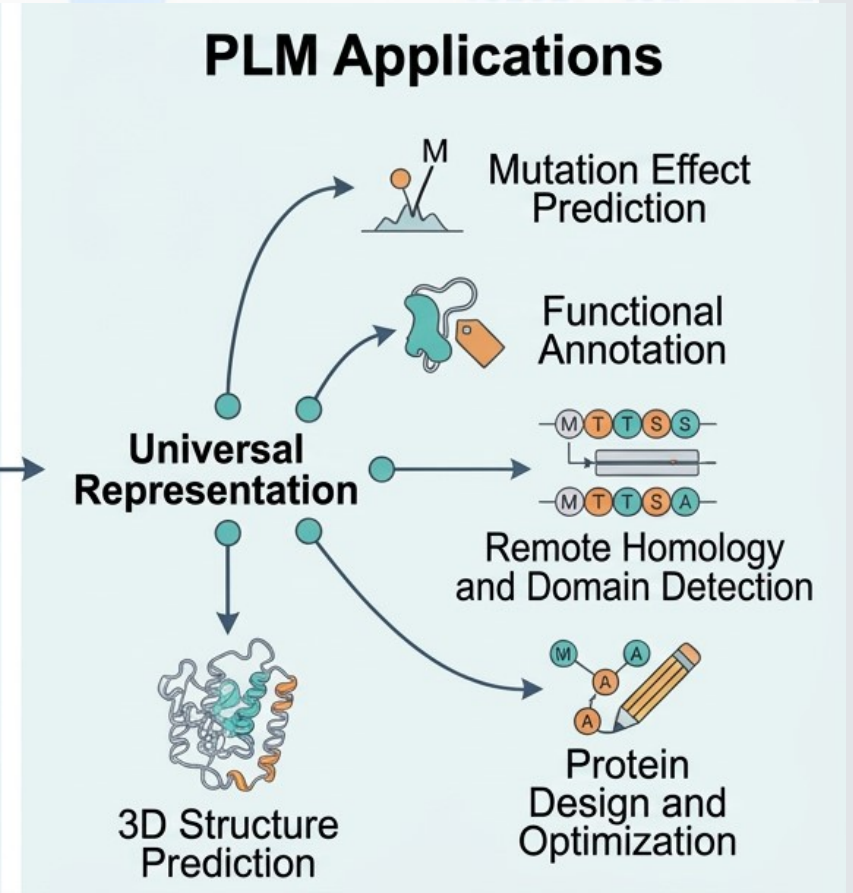
Common language model (LLM)



PLMs compress evolutionary, structural, and functional features into a single representation







ESM2 (Evolutionary Scale Modeling) — a state-of-the-art protein language model

Developed at Facebook* (FAIR); available on GitHub and Hugging Face

* Extremists and terrorists 🤦

Lin *et al.*, Science 2023

STRUCTURE PREDICTION

Evolutionary-scale prediction of atomic-level protein structure with a language model

Zeming Lin[1,2]†, Halil Akin[1]†, Roshan Rao[1]†, Brian Hie[1,3]†, Zhongkai Zhu[1], Wenting Lu[1], Nikita Smetanin[1], Robert Verkuil[1], Ori Kabeli[1], Yaniv Shmueli[1], Allan dos Santos Costa[4], Maryam Fazel-Zarandi[1], Tom Sercu[1], Salvatore Candido[1], Alexander Rives[1,2]*

Recent advances in machine learning have leveraged evolutionary information in multiple sequence alignments to predict protein structure. We demonstrate direct inference of full atomic-level protein structure from primary sequence using a large language model. As language models of protein sequences are scaled up to 15 billion parameters, an atomic-resolution picture of protein structure emerges in the learned representations. This results in an order-of-magnitude acceleration of high-resolution structure prediction, which enables large-scale structural characterization of metagenomic proteins. We apply this capability to construct the ESM Metagenomic Atlas by predicting structures for >617 million metagenomic protein sequences, including >225 million that are predicted with high confidence, which gives a view into the vast breadth and diversity of natural proteins.

The sequences of proteins at the scale of evolution contain an image of biological structure and function. The biological properties of a protein constrain the mutations to its sequence that are selected through evolution, recording biology into evolutionary patterns (*1*–*3*). Protein structure and function can therefore be inferred from the patterns in sequences (*4*, *5*). This insight has been central to progress in computational structure prediction starting from classical methods (*6*, *7*) through the introduction of deep learning (*8*–*11*) up to present high-accuracy structure prediction (*12*, *13*).

Language models have the potential to learn patterns in protein sequences across evolution. This idea motivates research on evolutionary-scale language models (*14*), in which basic models (*15*–*17*) learn representations that reflect aspects of the underlying biology and, with greater representational capacity, capture secondary structure (*14*, *18*) and tertiary structure (*19*–*21*) and show abilities such as few-shot language translation, commonsense reasoning, and mathematical problem solving, all without explicit supervision (*26*–*29*).

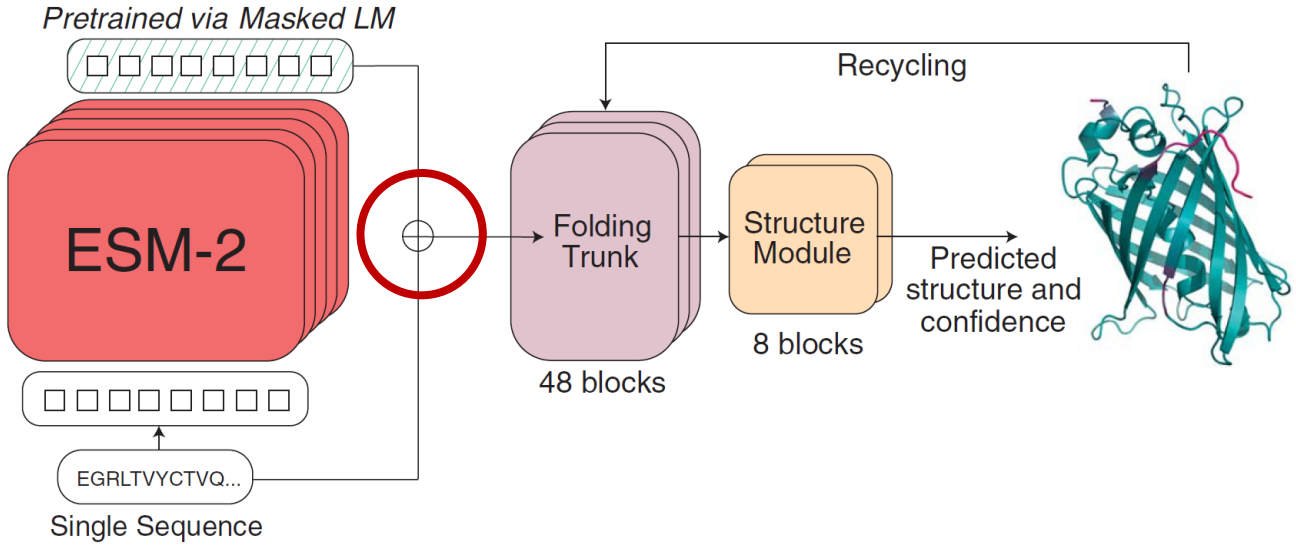
We posit that the task of filling in missing amino acids in protein sequences across evolution will require a language model to understand the underlying structure that creates the patterns in the sequences. As the representational capacity of the language model and the diversity of protein sequences seen in its training increase, we expect deep information about the biological properties of the protein sequences to emerge because those properties give rise to the patterns that are observed in the sequences. To study this kind of emergence, we scale language models from 8 million parameters up to 15 billion parameters. We discover that atomic-resolution structure emerges and continues to improve in language models over the four orders of magnitude in parameter scale. Strong correlations between

the high-sensitivity pipelines used by AlphaFold (*12*) and RoseTTAFold (*13*) and is a meaningful part of the computational cost even with recent lower-sensitivity fast pipelines (*30*). In practice, this means the speedup over the state-of-the-art prediction pipelines is up to one to two orders of magnitude.

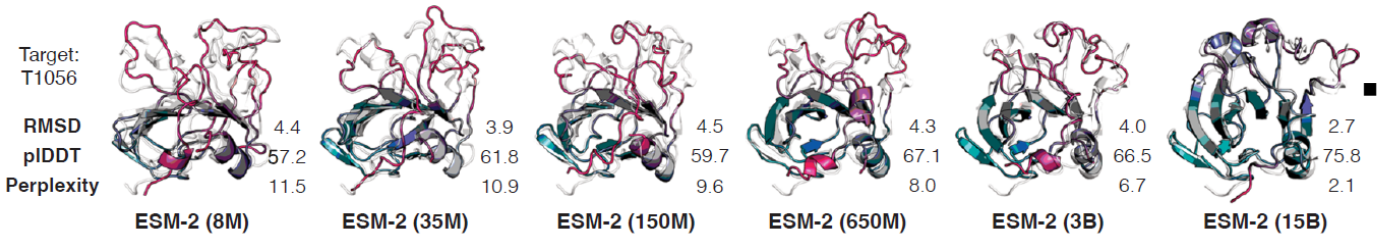
This speed advantage makes it possible to expand structure prediction to metagenomic scale datasets. The past decade has seen efforts to expand knowledge of protein sequences to the immense microbial natural diversity of Earth through metagenomic sampling. These efforts have contributed to an exponential growth in the size of protein sequence databases, which now contain billions of proteins (*31*–*33*). Computational structural characterizations have recently been completed for ~20,000 proteins in the human proteome (*34*) and the ~200 million cataloged proteins of Uniprot (*35*), but the vast scale of metagenomic proteins represents a far greater challenge for structural characterization. The extent and diversity of metagenomic structures is unknown and is a frontier for biological knowledge, as well as a potential source of discoveries for medicine and biotechnology (*36*–*38*).

We present an evolutionary-scale structural characterization of metagenomic proteins that folds practically all sequences in MGnify90 (*32*), >617 million proteins. We were able to complete this characterization in 2 weeks on a heterogeneous cluster of 2000 graphics processing units (GPUs), which demonstrates scalability to far larger databases. High-confidence predictions are made for >225 million structures, which reveals and characterizes regions of metagenomic space distant from existing knowledge. Most (76.8%) high-confidence predictions are separate from UniRef90 (*39*) by at least 90% sequence identity, and tens of millions of predictions (12.6%) do not have any match to experimentally determined structures. These results give a large-scale view into

ESM2 architecture



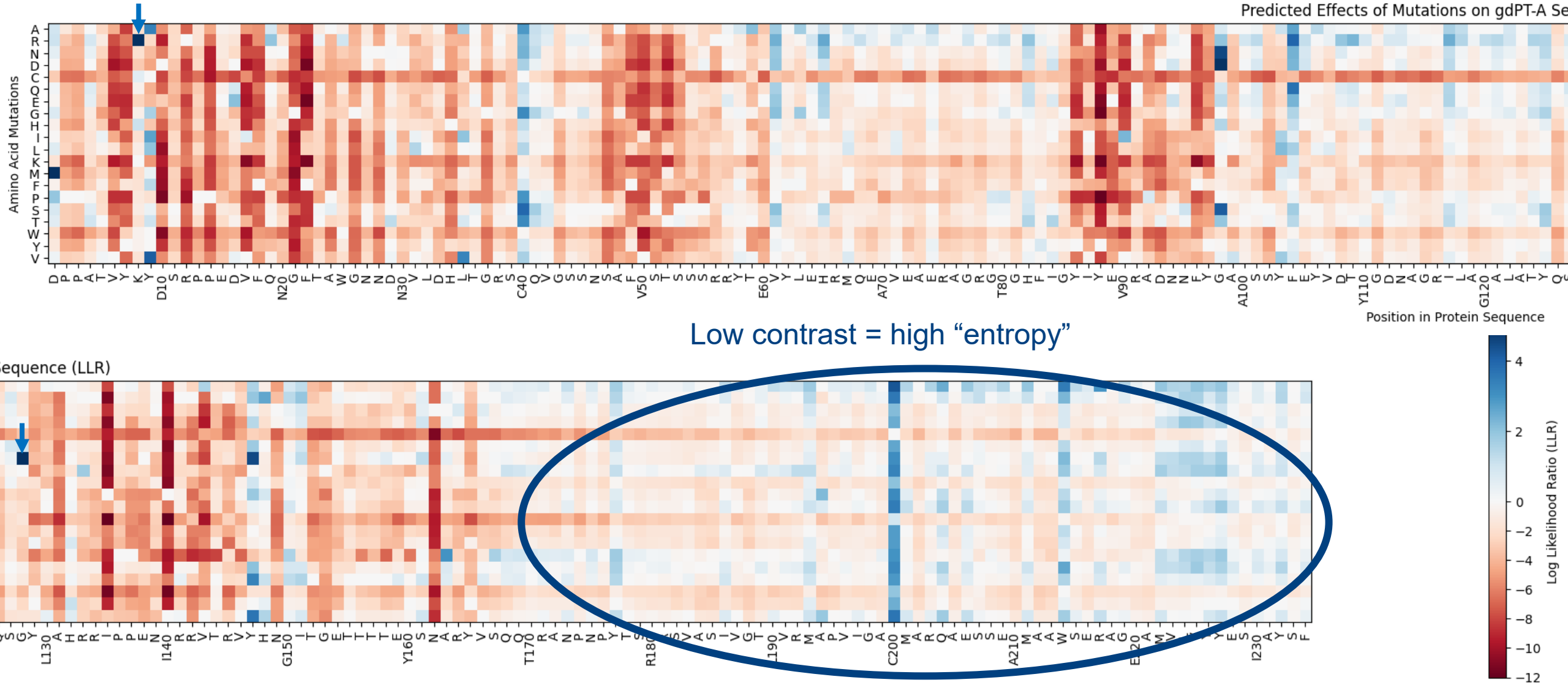
Model itself — Structure "head" (like in AF2)



Lin *et al.*, *Science* 2023

- BERT-style **transformer** (encoder-only, no decoder)
- Input: any amino acid sequence
- Outputs:
  - Embeddings — the basis for numerous ML applications
  - Logits (unnormalized token scores) — LLR maps and variant prediction
  - Contact maps — structure prediction with ESMFold
  - Attention maps — intraprotein patterns and key residues
- Training:
  - Masked language modeling (predicting 15% of masked amino acids)
  - Dataset: >250 million proteins (UniRef50)
  - No structures used in training — sequences only!
- Applications:
  - Screening mutations before synthesis (zero-shot prediction using LLR scores)
  - Predicting mutation pathogenicity
- Model sizes: 8M / 35M / 150M / 650M ("base") / 3B / 15B parameters
- Only the 15B model predicts 3D structures well (slightly less accurate than AlphaFold, but much faster)
- ESM Metagenomic Atlas (617M structures)

PT-A "tolerance map"



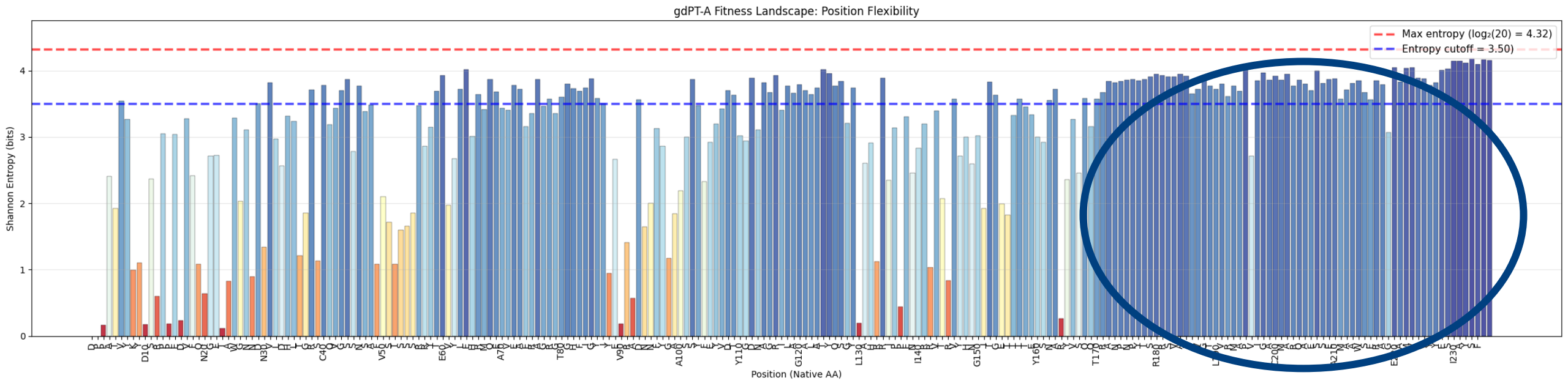
Overall **Redesign** Strategy

- Identify “hotspots” while freezing the rest
  - Mutational tolerance profile
  - Entropy map
  - ProteinMPNN ΔΔG
- Generate variant libraries ($N \times 10^4$)
  - PSSM (baseline)
  - ProteinMPNN
  - Inverse folding with homolog context (OpenProtein PoET2)
- Score using orthogonal methods
  - ESMFold
  - RMSD, pLDDT

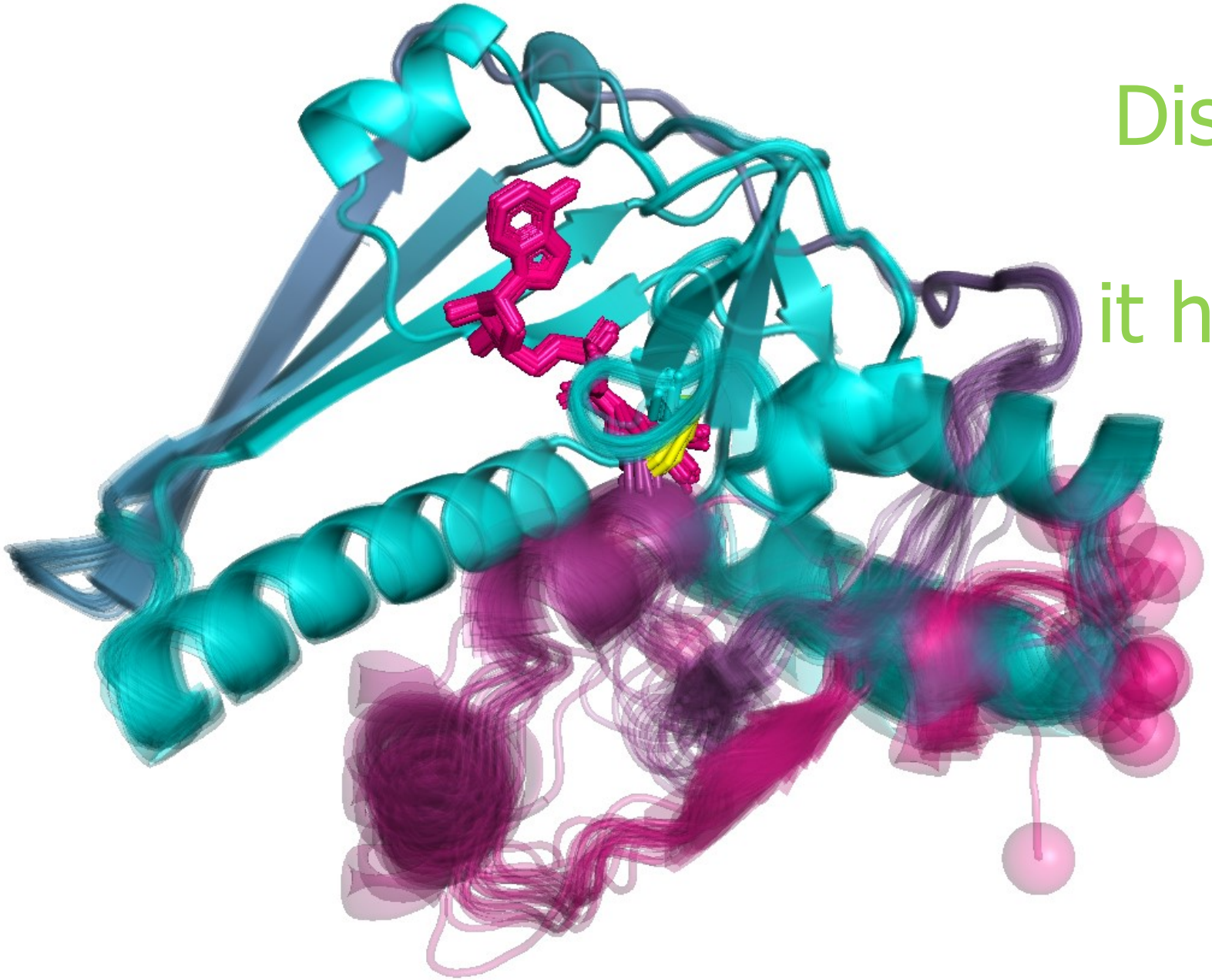


PT-A: two “domains”

- N-terminal catalytic (180 aa) — bound by neutralizing antibody 1B7
- C-terminus — “entropic”, disordered, hydrophobic:
  - Periplasm accumulation due to hydrophobicity
  - B-ring attachment
  - Unfolding leading to escape from endosome (ERAD)



Disorder of PT-A C-terminus: it has to be shorter!

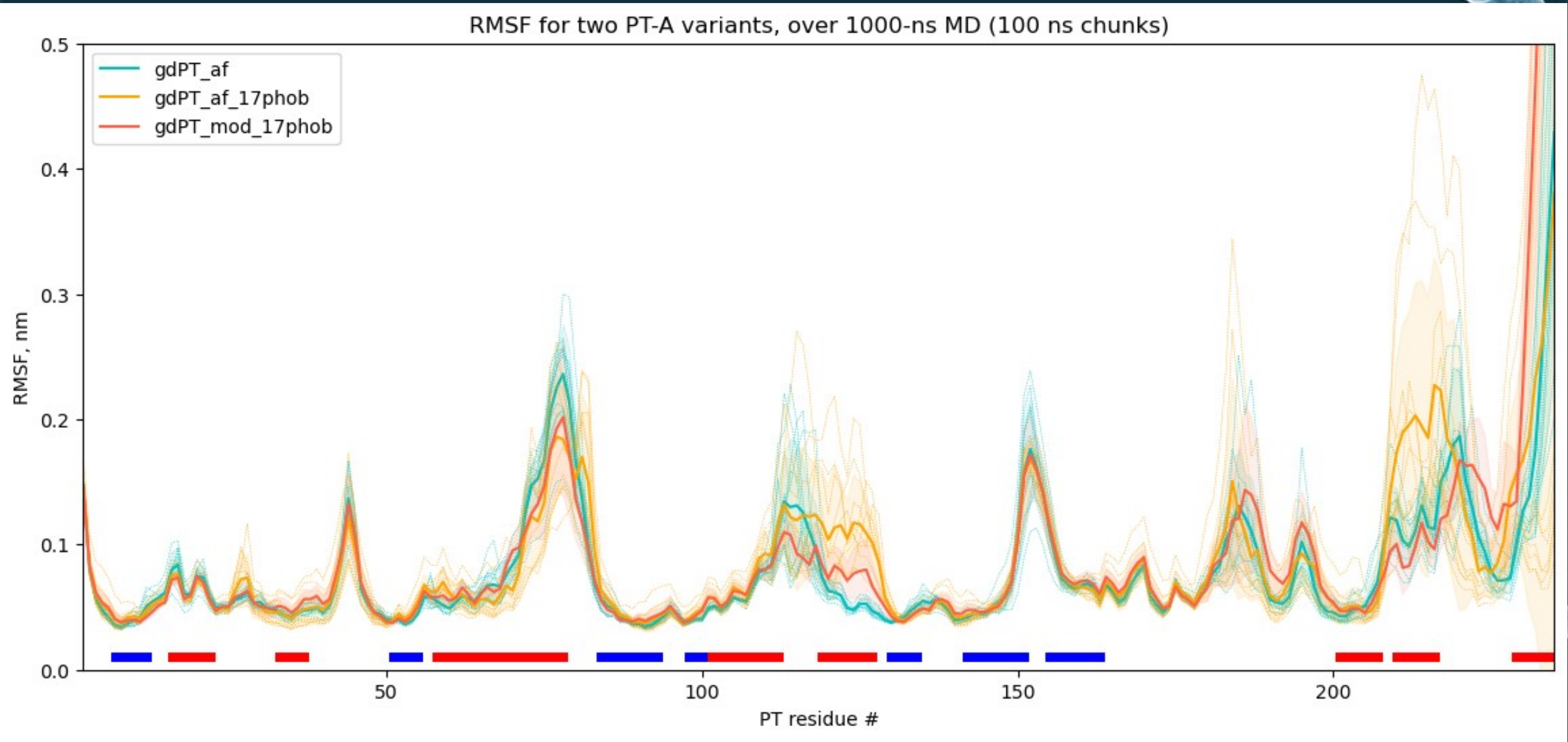


Boltz-2 cofolding:

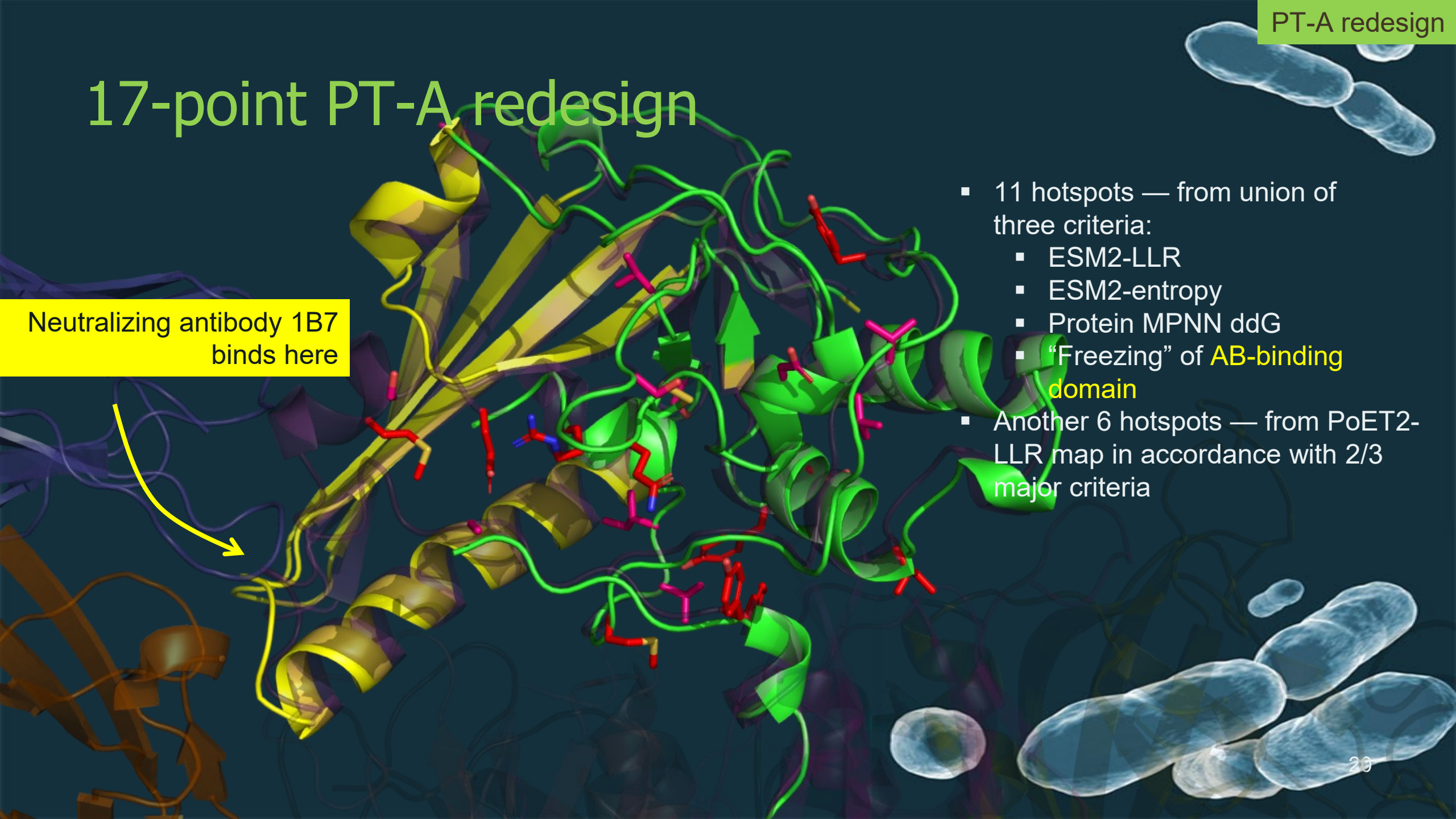
- PT-A + NAD
- No disulfide set

…but it seems unnecessary

MD of PT-A reveals the same

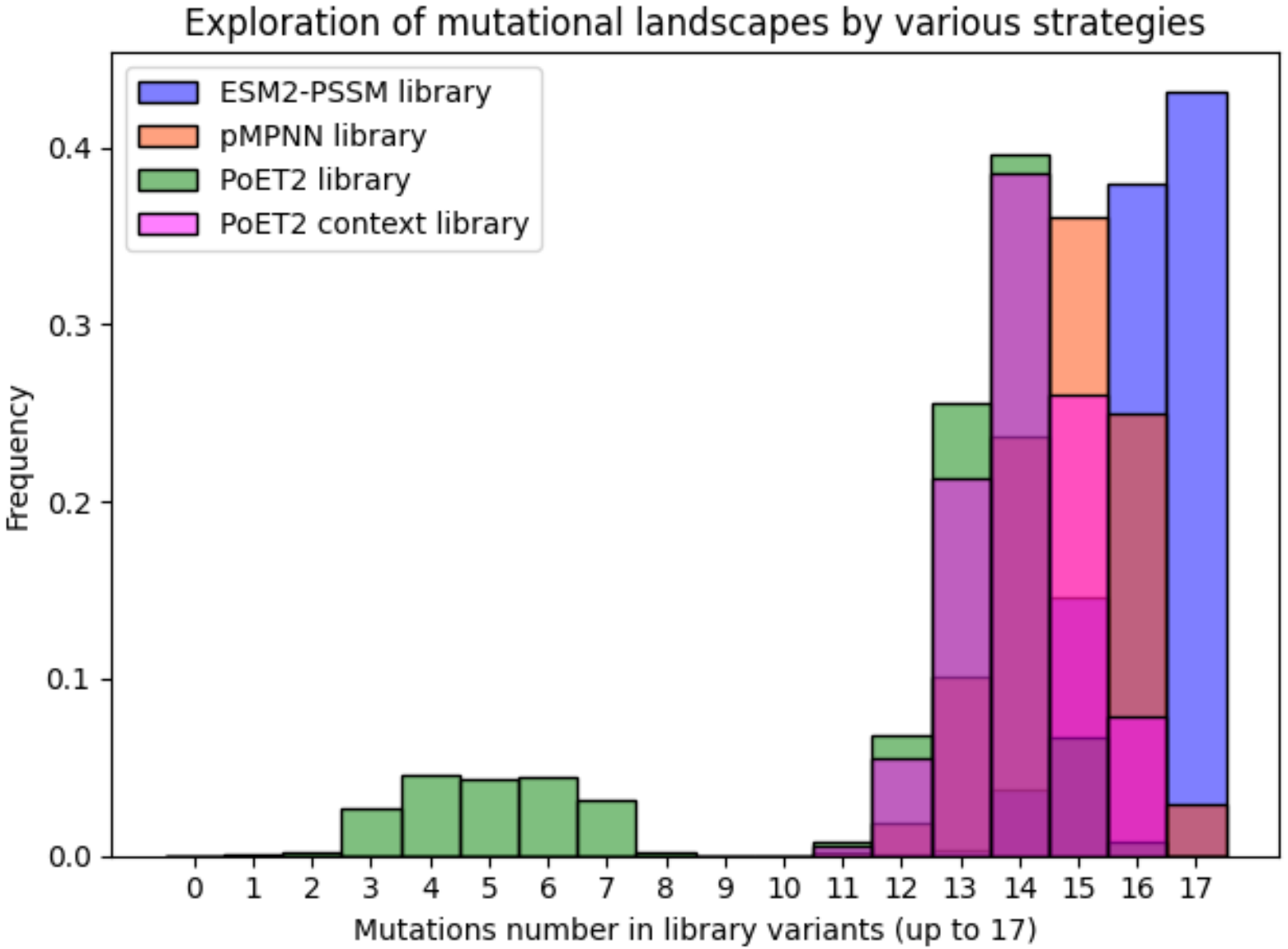


17-point PT-A redesign



- 11 hotspots — from union of three criteria:
  - ESM2-LLR
  - ESM2-entropy
  - Protein MPNN ddG
  - "Freezing" of AB-binding domain
- Another 6 hotspots — from PoET2-LLR map in accordance with 2/3 major criteria

17-point PT-A redesign



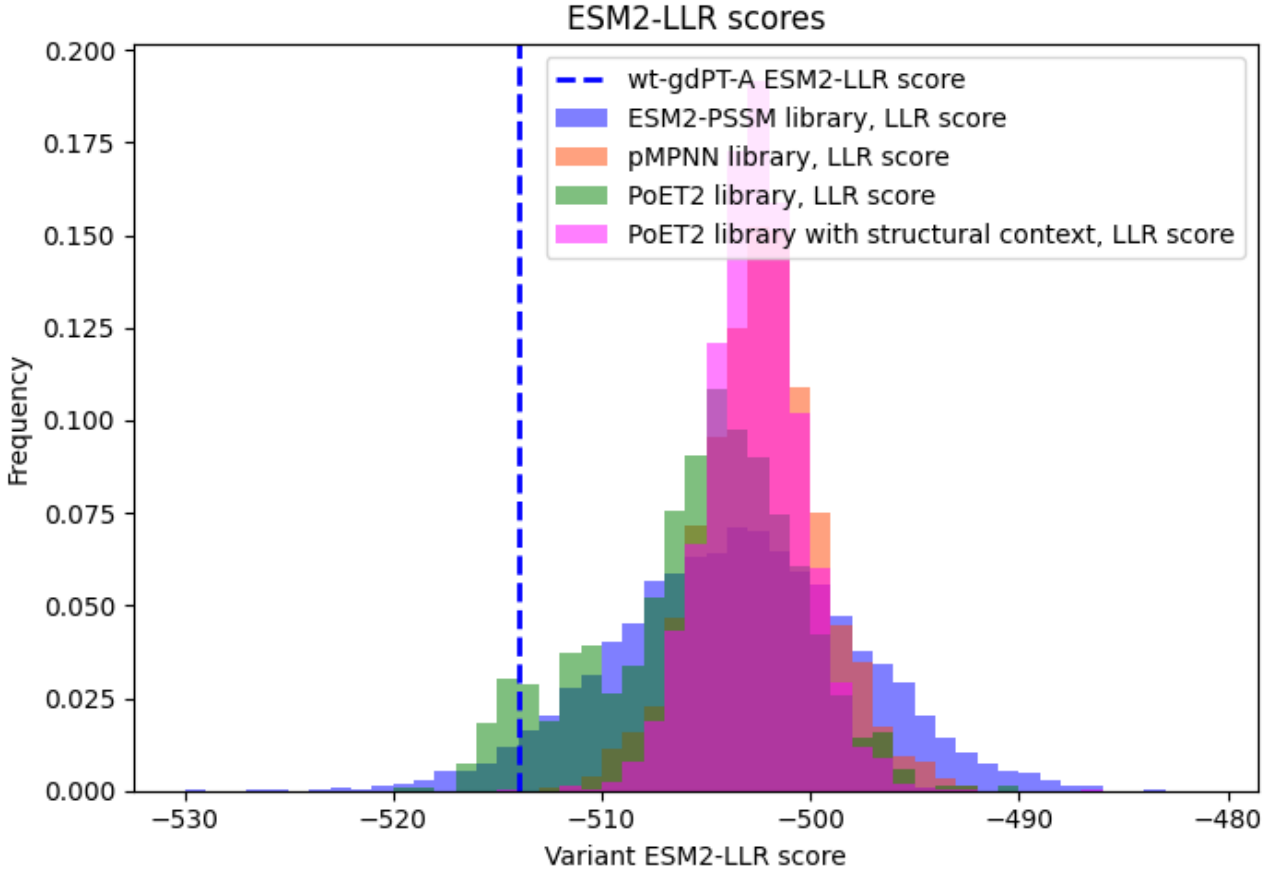
Four libraries sampled by different methods over 17 hotspots:

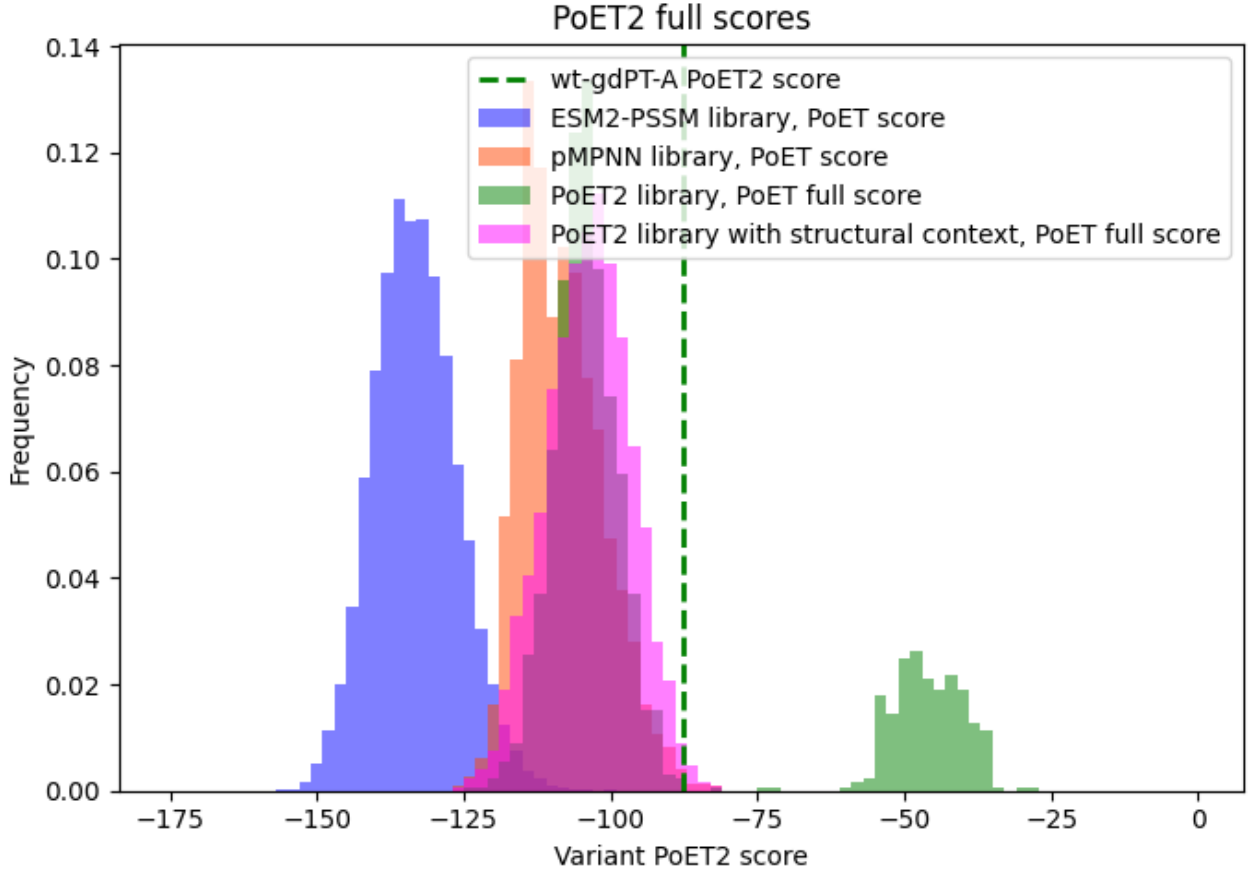
- PSSM (baseline)
- pMPNN: inverse-folding model
- PoET2: multimodal inverse-folding model employing structural and evolutionary context

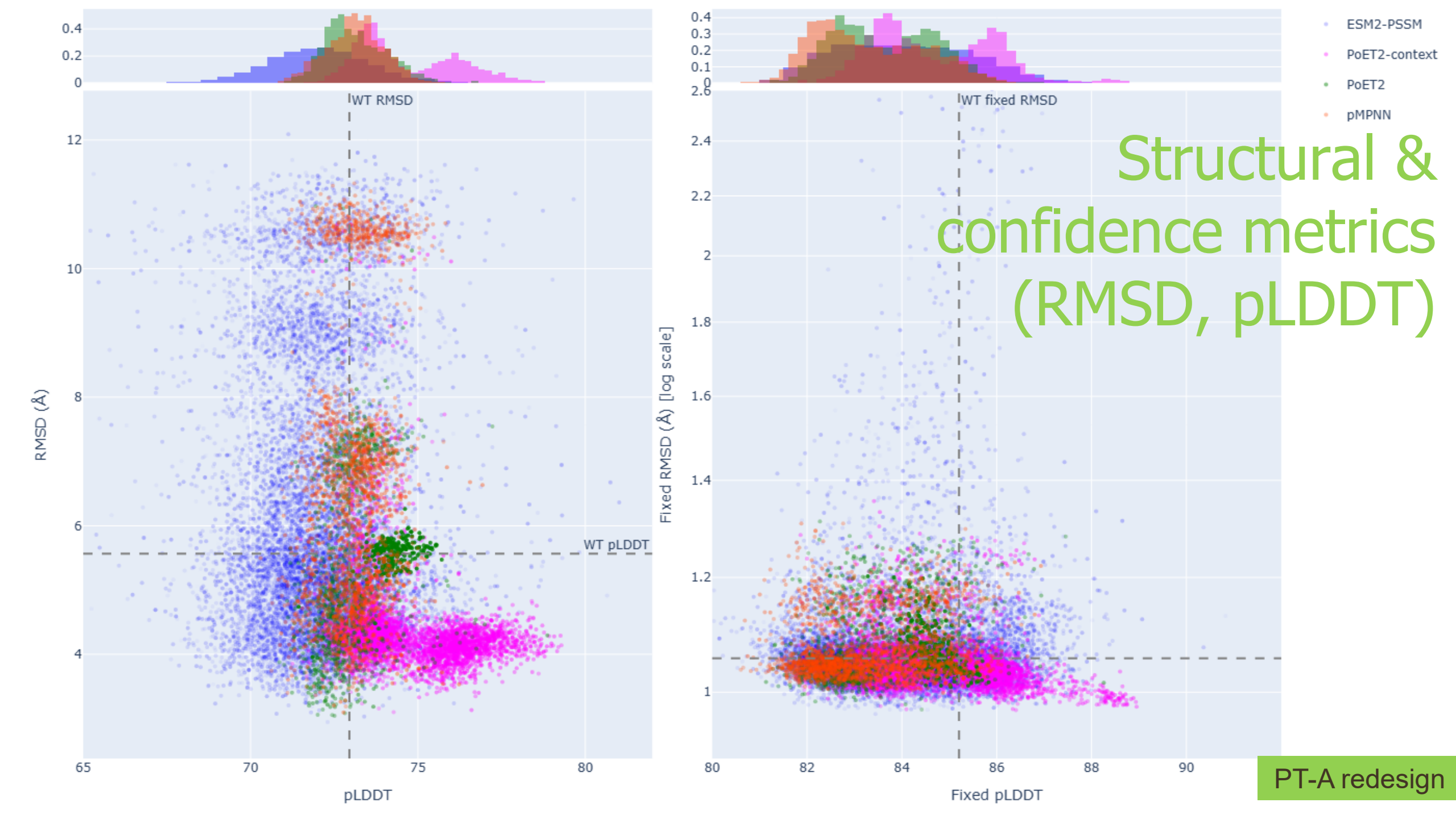
Understanding protein function with a multimodal retrieval-augmented foundation model
arXiv:2508.04724
August 2025

Sequences scoring

Generated libraries were scored by PLMs
Best: PoET2 with “right” context



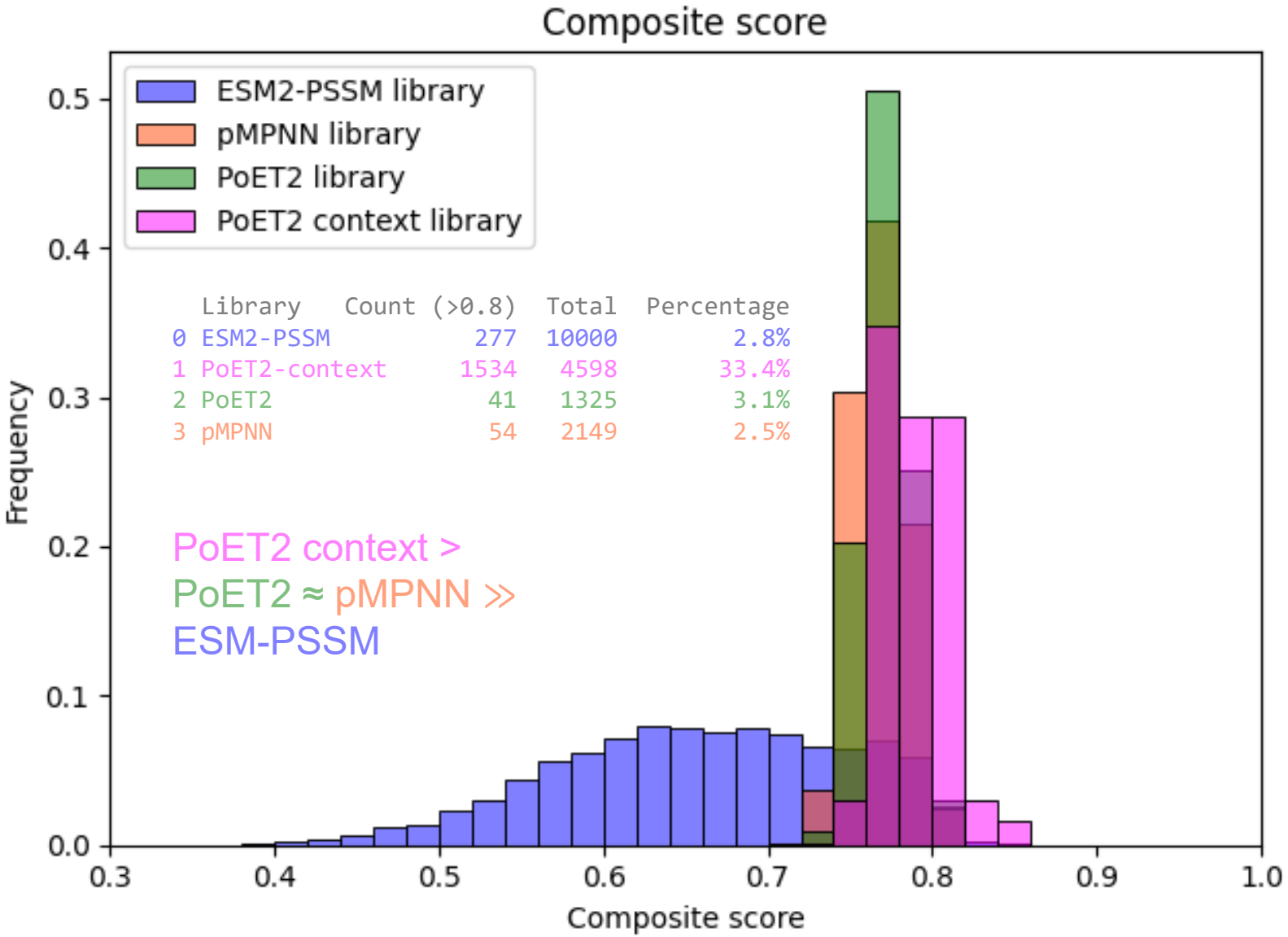




Structural & confidence metrics (RMSD, pLDDT)

PT-A redesign

Libraries ranking

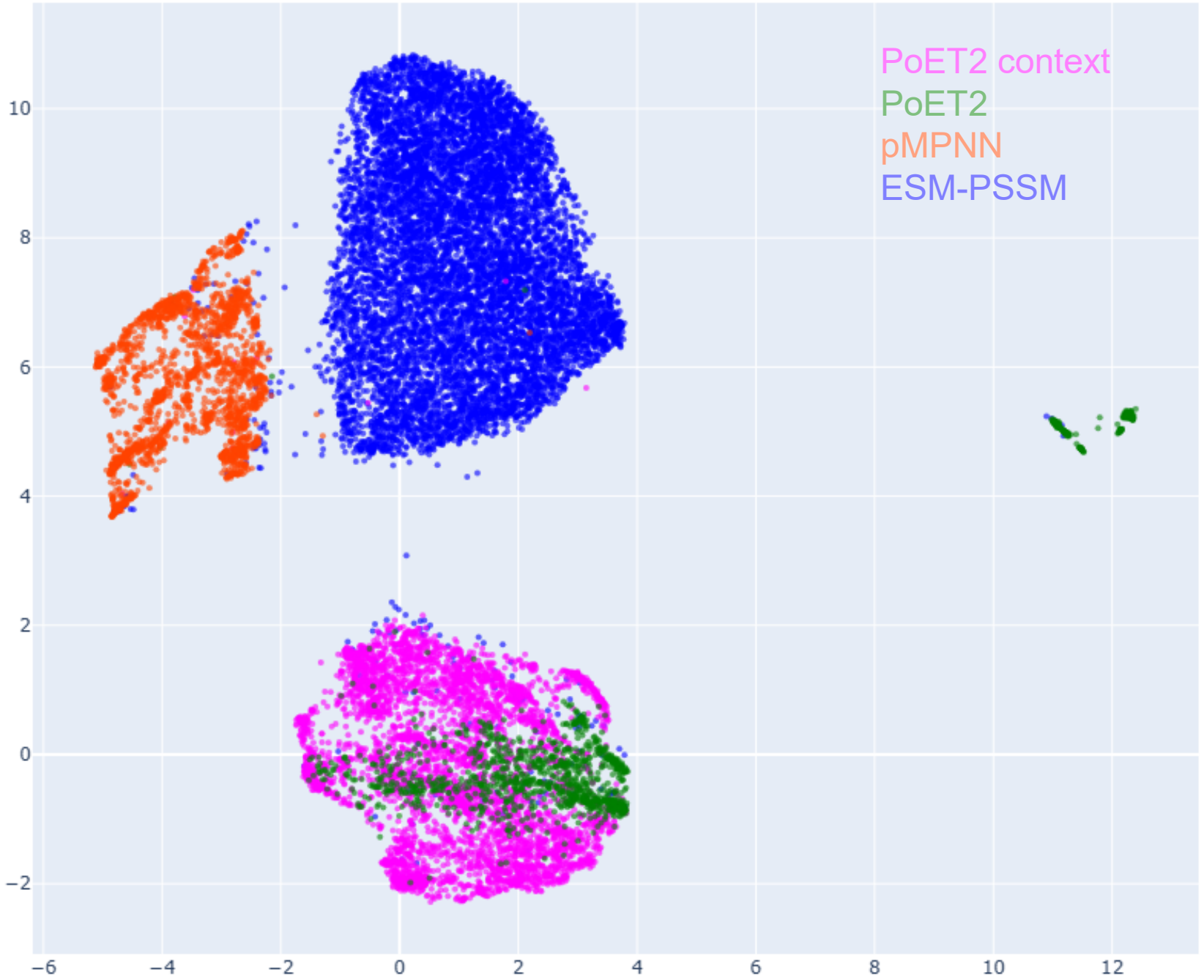


	Library	Count (>0.8)	Total	Percentage
0	ESM2-PSSM	277	10000	2.8%
1	PoET2-context	1534	4598	33.4%
2	PoET2	41	1325	3.1%
3	pMPNN	54	2149	2.5%

PoET2 context >
PoET2 ≈ pMPNN ≫
ESM-PSSM

Universal score combining (by Z-score):

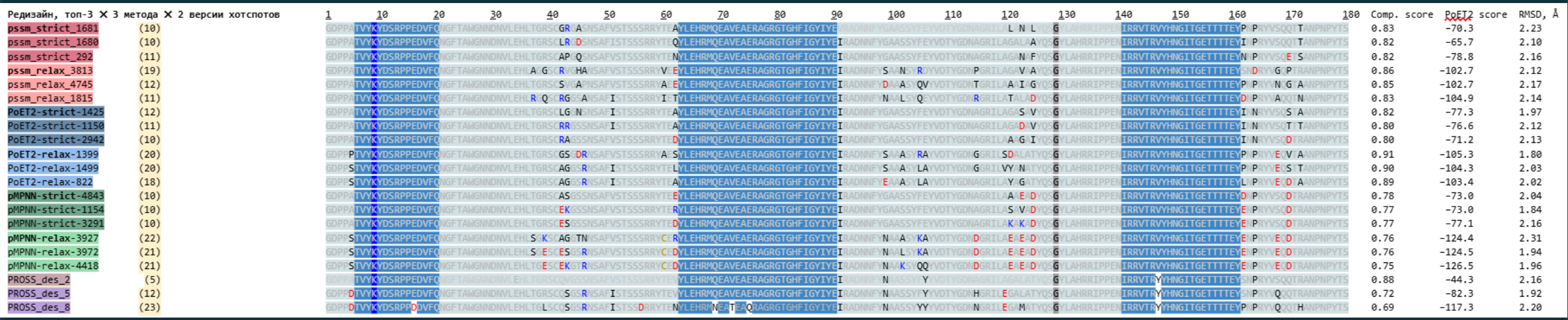
- RMSD
- pLDDT
- PoET2-score



Libraries visualisation

Uniform Manifold Approximation and Projection (UMAP) using ESM2-embeddings

Designed variants of gdPT-A_Δ55



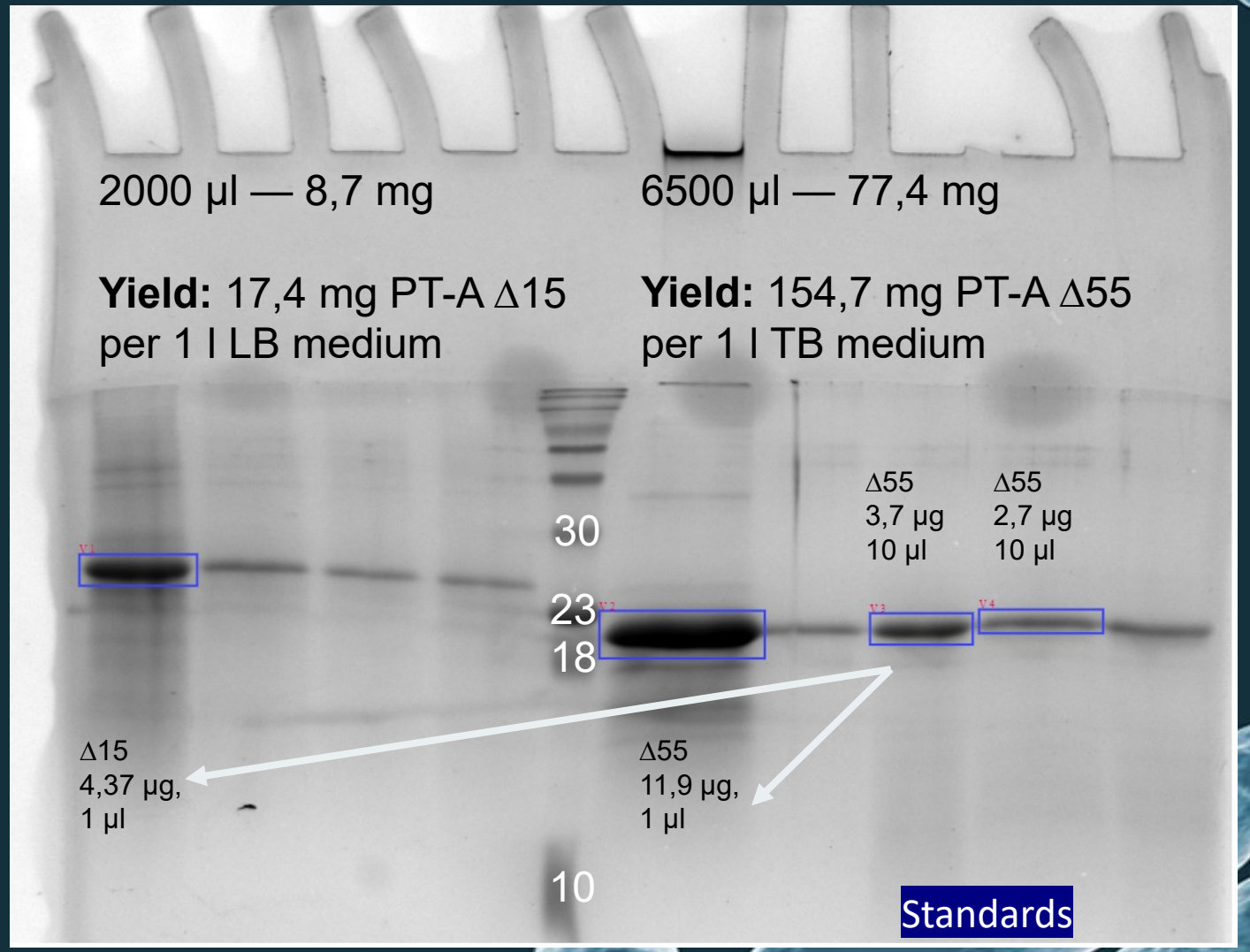
- Three methods / libraries
- Two hotspots versions
- Top-3 from each lib
- 21 variant to test

The only thing left — check if these proteins are more stable!

PT-A, in short!

Removal of the unordered C-terminus drastically increased yield and stability of antigen:

- PT-A: Δ15 → Δ55 (ESM2 hint)
- Medium: LB → TB
- *E. coli* strain: BL21 (DE3)
- Yield, mg/l: 17.4 → 154.7!



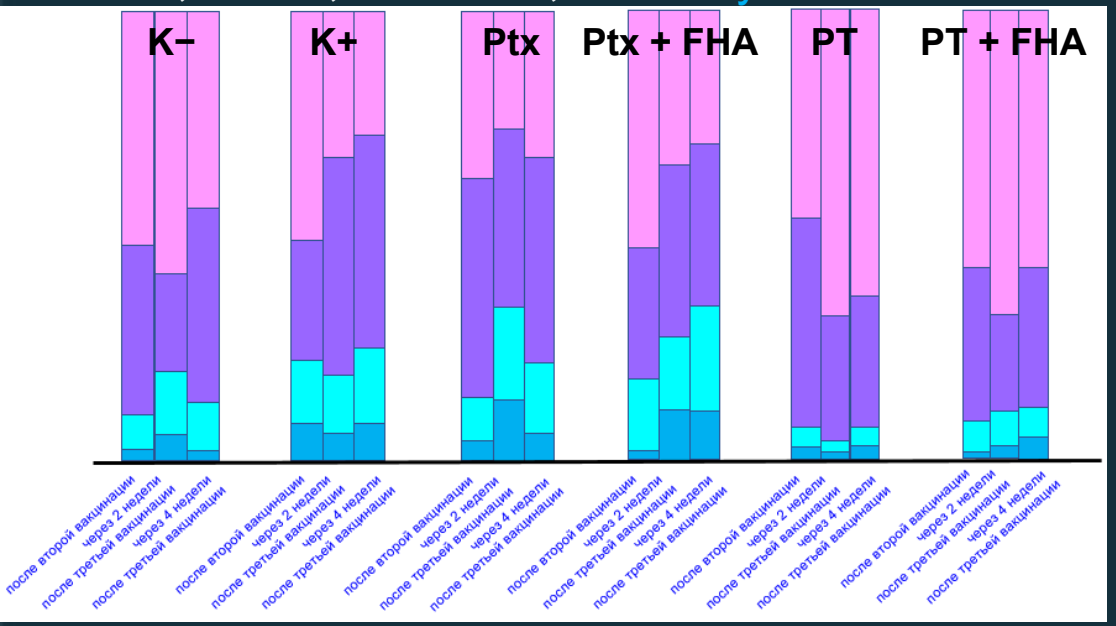
So far, only the effect of C-terminus removal has been confirmed, and experiments with 17-point mutants have not been conducted.

V. Matskevich results

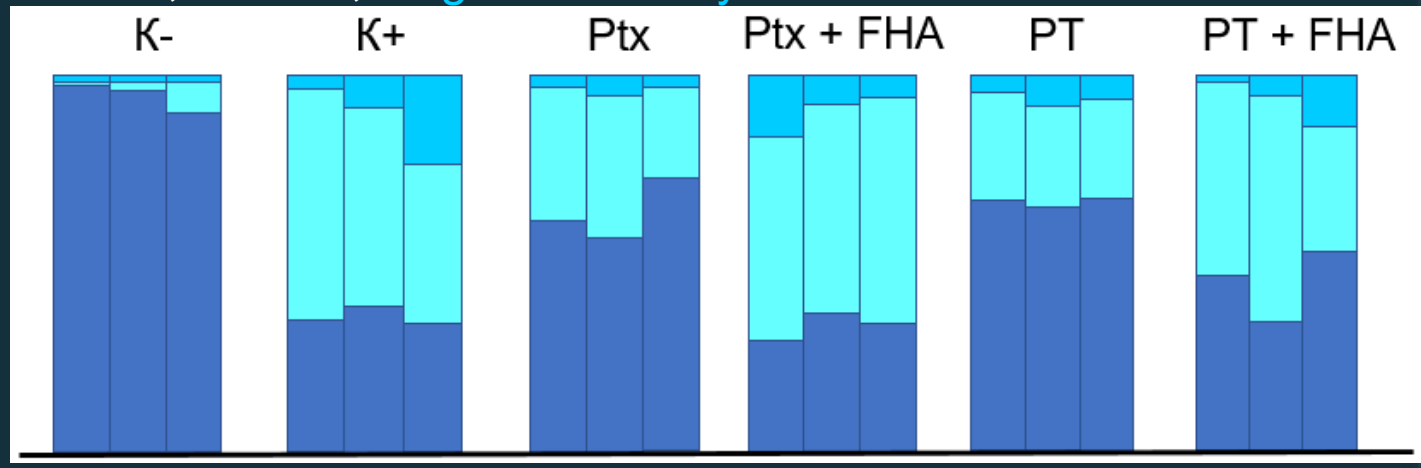
Results by A. Maleeva, V. Shokina

Cytometry: how the antigen works

Cytokine-producing cells:
CD3−, CD3+, CD4+ Th, CD8+ cytotox.



Memory cells:
naive, effector, long-term metory

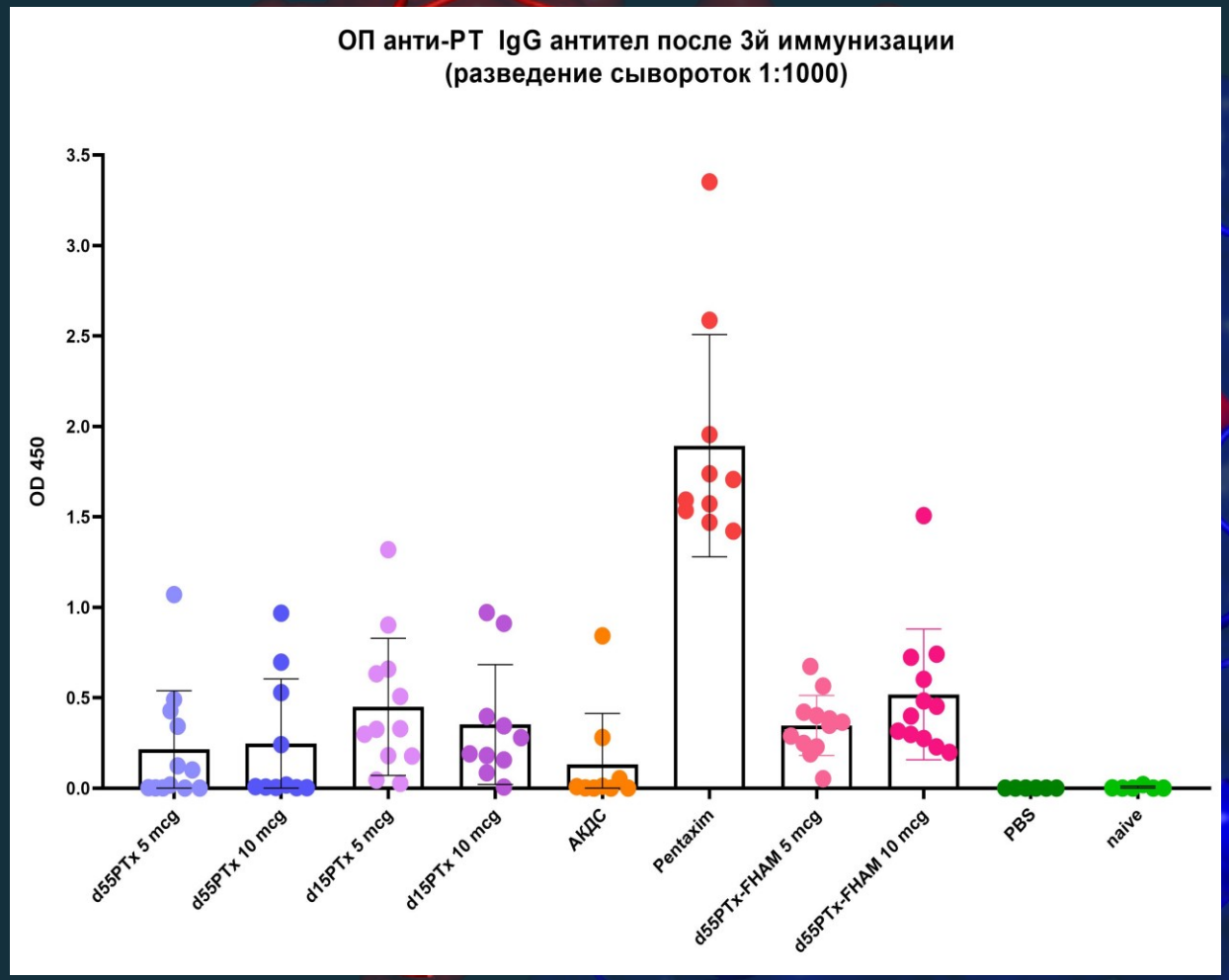


Animal groups : **K−:** negative control; **K+:** pentaxim;
Ptx: recombinant gdPT-A Δ15 (not Δ55!); **PT:** Full-size *B. Pertussis* toxin;
+FHA: addition of filamentous hemagglutinin (second subunit).
The experiments were conducted on cells from the spleen and lungs.

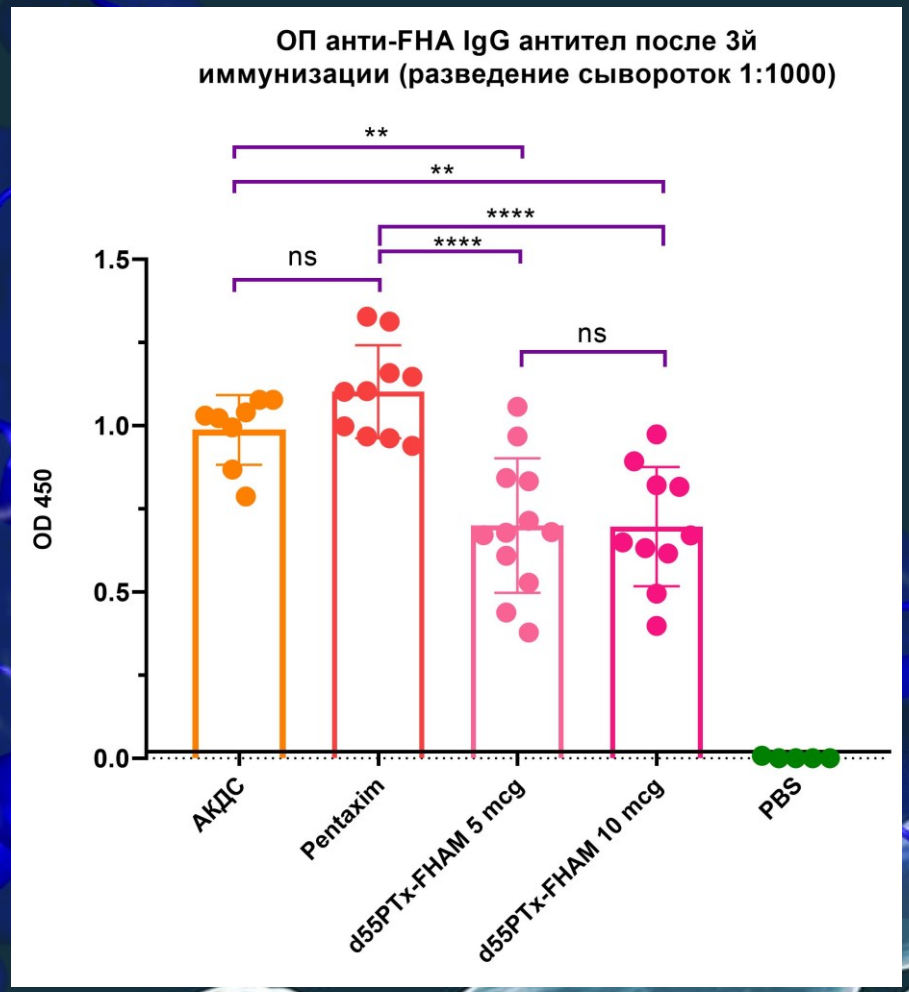
- Recombinant gdPT-A Δ15 elicits a stronger T-cell response than full-length PT, suggesting that Δ55 and its variants may perform even better
- Adding FHA enhances immunogenicity and improves T-cell memory
- Overall, however, T-cell memory formation is weak, with a bias toward effector cells — both for our formulations and for Pentaxim (positive control)
- This is a fundamental limitation of alum-adjuvanted subunit/acellular pertussis vaccines: more advanced **adjuvants** are needed

Results by Anna Lyapina

Next move: gdPT-A_Δ55—FHA chimera



Anti-PT antibodies are synthesized, although less compared to pentaxim



Anti-FHA antibodies are synthesized (although, less compared to controls)

Conclusions

- Recombinant gdPT-A Δ15 elicits a stronger T-cell response than full-length PT, suggesting that Δ55 may perform even better
- gdPT-A_Δ55—FHA chimera seems working, although with limited protection
- Overall T-cell memory formation is weak, with a bias toward effector cells — both for our formulations and for Pentaxim
- This is a fundamental limitation of alum-adjuvanted subunit/acellular pertussis vaccines: more advanced adjuvants are needed